May 2019

# A Simulation Platform for Generation of Synthetic Videos for Human Activity Recognition

Gary Plunkett
*Western Washinton University*

# Procedural Generation of Human Activity Videos for Training Action Recognition Systems

Gary Plunkett, Wesley Deneke, Emmanuel Harley, Sam Dixon
Western Washington University

## Objective

Create a simulation platform for outputting synthetic video of common household actions and workflows. Video is intended to be used to train neural network-based action recognition models. Procedural generation and randomness are introduced to increase video generalizability.

**Figure 1.** Example actions in a generated household workflow. A human agent opens the fridge, picks up the beer, closes the fridge, then sits on the couch and turns on the TV.

## Motivation

**Labelling video data is expensive!**

- Modern computer vision systems are trained on large data sets of real video.
- Need to draw object boundaries on each frame. Known as semantic segmentation (fig 2).

**By generating our own synthetic video data, we get labels for free.**

- Automatically output semantic segmentation, and other useful ground truth output (fig 3).
- Training with synthetic and real video has been shown to improve performance of computer vision system. [1]

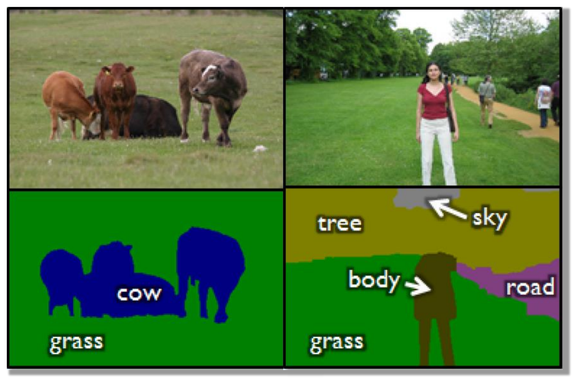


**Figure 2.** Examples of semantic segmentation. Segmentation for every frame should be provided to optimize action recognition network training.

## Design Criteria

- Produce a simulated environment of a residential home with high photorealism using Unity.
- Ensure simulated actions can only be performed by agents in a physically plausible manner.
- Represent complex activities as a hierarchy of workflows composed of actions.
- Automate dimensions of variation for video diversity, including object placement, character models, lighting, camera angle, and action sequencing.
- Output synthetic videos in multiple modalities for each ground-truth annotation (fig 3).

**Figure 3.** Synthetic video and accompanying ground truth for opening a refrigerator: RGB, Normals, Optical Flow, Depth, Semantic Segmentation.

## Implementation

**Workflow Specification:** Manages the action execution order, enforcing conditional and branching execution paths. Interactable objects are specified in the abstract.

**Simulation Generator:** Translates workflow specification to specific object instances. Introduces variation in lighting, camera angle, object locations, object type, room layout, human model, and agent animation parameters.

**Interaction System:** Backend system for controlling interactions between agent and household objects. Handles change in game state, object animations, and inventory.

**Inverse Kinematics:** Inverse kinematics are used to create plausible agent-object interaction animations. The IK system generates an interaction animation based on the agent's current position, and interaction hand end-point.

**Video Data Pipeline:** Video data from multiple modalities is saved, along with current atomic action and current workflow.

## Evaluation

Our simulation platform is designed to facilitate research into workflow recognition systems, where a workflow consists of multiple atomic actions. Before we can investigate workflow recognition, we must validate our synthetic data's usability by training action recognition systems on one action at a time.

**Model:** We will use a Multi-Stream Inflated 3D-ConvNet (I3D), an action recognition neural network (fig 4). This model reports state-of-the-art results when trained on multiple inputs streams. [2]

**Video Data:** Our initial classification task will be kept as simple as possible. The model will be trained to distinguish between a "PickUp" action, and an "Open" action. The streams of video input to be shown in figure 3, with the addition of Instance Segmentation data as well.

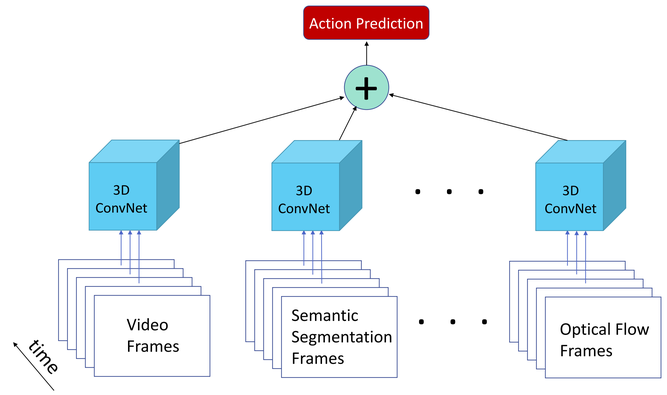


**Figure 4.** Schematic of the I3D action recognition network to be used for synthetic video evaluation.

## Future Work

**Workflow recognition**, the identification of a string of atomic actions, is an almost completely unexplored area in machine learning. A purely linguistic approach could be layered on top an action recognition network, or multiple recognition networks could operate at different frequencies a la SampleRNN. [3]

**Expansion of the simulation platform** is another large priority. Currently, all actions are confined to the kitchen and the living room of a home. Adding interactable functionality to the rest of the house, or adding a new home altogether, would increase the diversity of generated video.

## References

1. C. De Souza, A. Gaidon, Y. Cabon, and A. Peña, "Procedural Generation of Videos to Train Deep Action Recognition Networks," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2594-2604, Jul 2017.
2. Joao Carreira, Andrew Zisserman. "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," https://arxiv.org/abs/1705.07750, May 2017.
3. Mehri et al. "SampleRNN: An Unconditional End-to-End Neural Audio Generation Model," *ICLR 2017 Conference,* https://arxiv.org/abs/1612.07837, Feb 2017.